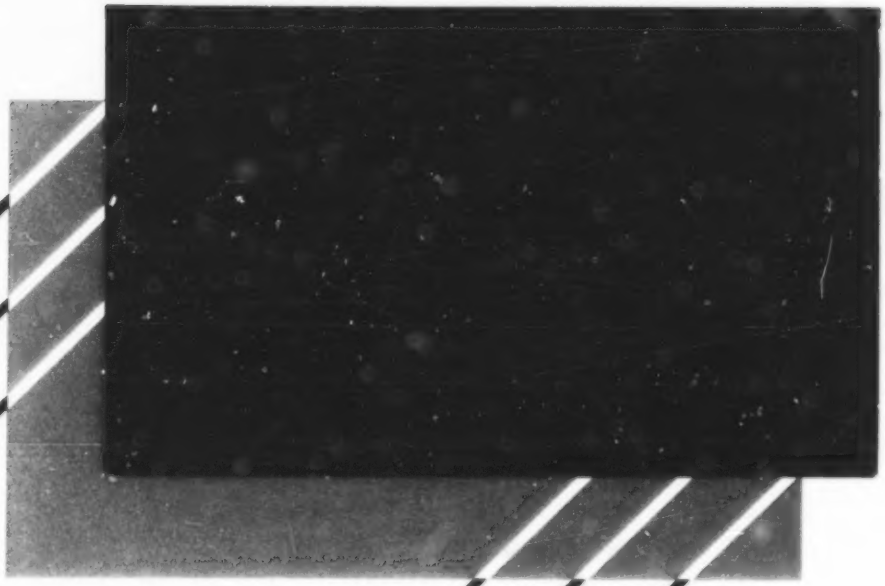
Gouvernement du Canada
Ministère des Communications

Government of Canada
Department of Communications

Le Centre canadien de recherche sur l'informatisation du travail
Canadian Workplace Automation Research Centre



Canada

MACHINE-AIDED TRANSLATION:

STATE OF THE ART AND PROSPECTS FOR THE FUTURE

Pierre Isabelle

Canadian Workplace Automation
Research Centre (CWARC)

Copyright Minister of Supply and Services Canada 1988.

Cat. no. Co 28-1/19-1988 E

ISBN 0-662-16397-4

The views expressed in this report are those of the author only.

* Rapport complet en français aussi disponible.

The purpose of this article is to present the CWARC's machine-aided translation research and development program by situating it in the context of an assessment of the current state of the art in computerized translation.

THE PROBLEM OF AUTOMATING TRANSLATION

Even since the inception of the digital computer in the fifties, many people have believed that the emergence of a universal translating machine was imminent. However, a cursory examination of the current situation shows that the past thirty years have produced meagre results. Most translators working today are still using methods little different from those of a century ago.

Over the years, the initial optimism with which machine translation was greeted has given way to grave doubts and even to total pessimism among informed observers. People began to realize that translation between natural languages is infinitely more complex than a simple transcoding process.

The principal reason for this is that the information explicitly given in any text leaves a great deal of indeterminacy as to its overall informational content. As early as 1960, Bar-Hillel (who had previously been highly optimistic for the future of machine translation) put forward the thesis that fully automatic translation could never produce high-quality results.[1] Bar-Hillel's argument rested on examples such as the following:

(1) The box is in the pen.

A bilingual dictionary will tell us that "pen" can sometimes be translated as "plume" and sometimes as "parc" (as in "playpen"). The correct choice for translating Bar-Hillel's example depends not on the explicit information contained in the text but rather on our knowledge of the world (e.g., the relative sizes of objects, the relative probability of various situations). It is quite easy to find other examples where correct translation presupposes extralinguistic knowledge of the real world:

(2) ... federal and provincial governments.
(3) John dropped the glass on the table. It broke.
(4) Discard used filter and replace it.

In example (2), translation into French requires the selection of agreement for number in the case of each adjective: the choice of "fédéral (singular) et provinciaux (plural)" derives from the translator's knowledge of a certain political reality. To translate the pronoun "it" correctly in (3) -- "il" (masculine) or "elle" (feminine) -- one must know whether it was the glass or the table that broke ("le verre", masculine, or "la table", feminine); in the absence of any indication to the contrary, our common sense leads us to prefer the masculine pronoun "il". Lastly, example (4) requires the translator to choose between expressions meaning either "put back in its place" or "substitute" for the verb "replace". Here again, we rely on common sense: when a disposable item is thrown out, it is not normally retrieved and immediately put back in service.

The inescapable conclusion that emerges from such examples is that fully automatic high-quality translation of arbitrarily selected texts is possible only to the extent that we have intelligent machines; i.e., machines capable of accessing not only the explicit linguistic content of texts (morphology, syntax, semantics) but also their implicit content, derived from a generalized cognitive system that includes common sense, encyclopedic knowledge, reasoning, etc.

If, over the last 15 years, the ambitious venture now known as artificial intelligence has made any tangible progress, it has given us better insight into the daunting complexity of the human cognitive apparatus. We are still very far from creating a universal thinking machine. Consequently, we are still far from having a universal translating machine.

THE SEARCH FOR SUB-OPTIMAL SOLUTIONS

The worldwide translation market has been experiencing sustained expansion in recent years. The economic and political stakes involved are growing inexorably.

If the optimum solution (the universal translation machine) cannot be found, then it is important to find remedial solutions to help translators cope with demand that is becoming ever harder to meet at a reasonable cost. These solutions will be based on compromises over one or more of the following attributes of a universal translation machine: (a) ability to produce high-quality translations; (b) full automation; (c) ability to translate any given text.

(A) Approximate translation

It sometimes happens that specialists are required to scan masses of documentation in a foreign language in search of a small relevant sub-set of data. Machines capable of producing very rough drafts may prove useful, even if the parts identified as relevant then have to be translated manually. Situations where this approach is useful (e.g., military intelligence) are the exception rather than the rule, however.

(B) Semi-automatic translation

There is a range of more widely applicable techniques based on partial automation of the translation process. In the case of machine-aided human translation, the human translator retains the initiative for the translation process but has access to a machine that can automatically perform certain peripheral tasks. Functions that translators could find useful might include the telecommunication of texts, access to lexicographical and terminological databases, access to banks of existing translations, multilingual word processing, spelling checkers, etc. This technologically modest approach has the advantage of being suited to almost all translation contexts. We can only endorse Kay's[2] complaint that too few people have hitherto taken advantage of its potential.

Human-assisted machine translation differs from the foregoing approach in that the machine is entrusted with the primary initiative for the translation task but with human assistance available to solve problems beyond its capacity. This human intervention may precede the automatic translation process, in which case annotations are made to the text; we then speak of systems based on human **pre-editing** (e.g., the University of Hong Kong's CULT system). The intervention may also take the form of information provided to the machine on demand; such systems are called **interactive** (e.g., the ALPS system). Finally, intervention may come after the fact, in the form of correcting errors and filling in gaps, known as **post-editing**. Most machine translation systems require some post-editing. In contrast, pre-editing and interactive systems present some thorny problems: the information the machine requires often has to be provided in a form the machine can readily handle but which is not at all natural for a human.

(C) Specialized machine translation

If both high-quality and relatively advanced automation are required, then "universality" can be waived in favour of tailor-made systems designed for highly specific types of texts. This category includes, first, **restricted input systems**; i.e., systems which presuppose that the material to be translated has been drafted in simplified language, in accordance with a rigid set of rules. The TITUS system, which translates abstracts of articles in the field of textiles, is the best-known example. This approach can be used only where translators have control over the drafting process and the subject matter is readily reducible to highly codified language.

Specialized systems can also be developed for natural sublanguages.[3] This approach, developed by the University of Montreal's TAUM group[4], relies on the existence of language sub-systems which evolve spontaneously within linguistic sub-communities in certain situations. Weather forecasts, stock market reports and recipes are all typical examples of natural sublanguages. In each case, pragmatic constraints on text content combine with very tight restrictions on the language in which the content is cast. The TAUM-METEO system has been translating Environment Canada's daily weather bulletins since 1977, and its workload is now approaching 15 million words per year. Its post-editing requirements are minimal, and this system is undisputably the most spectacular success in machine translation to date.

Thus, automation of translation may be viable and highly cost effective, in so far as we can implement appropriate sub-optimal techniques.

RENEWED INTEREST IN MACHINE TRANSLATION

The initial optimism surrounding the possibility of developing a translating machine generated intense research and development activity during the fifties. At that time, machine translation held pride of place in the realm of "the mechanization of thought processes", or what today is called "artificial intelligence".

In 1966, a committee mandated by the U.S. government to look into the causes of repeated failures in machine-translation projects submitted its report, now known as the "ALPAC Report".[5] Its authors were severely critical of the short-term approach characterizing the majority of projects and recommended that government funds be redirected toward longer-term research in computational linguistics and artificial intelligence. This report had a devastating effect on the field of machine translation, wiping out most R&D projects worldwide.

A careful reading of the reports, however, reveals that it did not criticize machine translation as a field of study, but rather short-range approaches in artificial intelligence. Its effect, though, was to discredit the very notion of machine translation.

Even today, research in machine translation is practically non-existent in the USA. Many researchers in computational linguistics still prefer to focus their efforts on applications that address much less-obvious needs than those of translation.

Since 1980, there has been a resurgence of interest in machine translation. In the United States, this movement is particularly evident in the private sector. There are now three new commercial systems (ALPS, WEIDNER and LOGOS), as well as SYSTRAN, which had enjoyed a near monopoly since 1965, and there would appear to be many takers for these systems in a number of countries. In addition, IBM and Digital Equipement are now pursuing work on machine and machine-aided translation.

In Europe, the EEC has launched a large-scale initiative with the ambitious EUROTRA project, whose goal is to construct a system capable of translating among all nine languages of the Community. Several European countries are funding their own university research programs (France, Holland, Britain). Siemens (Germany) and Philips (Holland) are also developing their own translation systems.

The most intense activity, however, is in Japan, where there are no fewer than 15 university and corporate R&D projects. Furthermore, machine translation is at the centre of the second phase (now getting started) of the well known fifth-generation computer project.

CWARC'S RESEARCH AND DEVELOPMENT PROGRAM

The Canadian Workplace Automation Research Centre is currently setting up a research and development program in machine-aided translation. This program was inspired by the recommendations of the COGNOS report[6] and reflects the view of the state of the art presented in the foregoing sections. Accordingly, two of the program's three projects are intended to develop systems based on specific sub-optimal techniques. The third part of our program concerns research on more powerful translation systems, based on artificial intelligence techniques.

Project I: Translator's workstation

The only widely applicable systems which may currently be developed for translation are those in which the human translator retains the entire initiative for the translation process, providing an integrated set of tools designed to help in this task.

The translator's work is complex and has many different facets. As a result, a specialized translator's workstation should include and properly integrate a wide range of tools, among which should be the following:

- electronic text transmission;
- translation management support;
- lexicographical and terminological research support
  - a computerized Kardex
  - computerized natural-language dictionaries
  - access to remote terminology banks
  - access to banks of existing translations
  - tools for setting up concordances;
- Writing and editing support
  - multilingual word-processing
  - spelling and grammar checkers, etc.

Lastly, this type of station should allow for the later integration of more advanced modules; e.g., specialized machine translation systems.

Project II: Specialized machine translation

Earlier, we saw that the so-called sublanguage approach has had spectacular results in the field of weather forecasting (TAUM-METEO system). We feel that it is extremely important to capitalize on this type of technique for as wide a range of subject areas as possible. Whereas the workstation which is the focus of Project I opts for generality of application at the expense of automation, this project does the opposite. The two approaches are perfectly complementary since, in the medium term, the workstation will provide an ideal environment for implementing specialized machine translation modules.

Before starting to develop a specialized translation system, it is essential to get an exact measure of the complexity of the sublanguage in question. This will involve listing all possible areas of application and evaluating their relative complexity, using the best metrological tools available.

Only then can we proceed with one or more development projects, using proven programming tools and techniques for natural language processing.

Project III: "Third generation" translation systems

To the extent that the projects described above meet their objectives, our program will have made a tangible medium-term contribution to the solution of translation problems. It has to be acknowledged, however, that

this contribution will be quite modest: the workstation only automates tasks peripheral to the translation process, and the specialized systems only address the particular case of simple sublanguages.

If we hope to extend beyond these limits, it is absolutely necessary to promote research directed toward more intelligent systems. Although research on natural-language processing and artificial intelligence has received considerable support in recent years, the fact remains that this research has rarely been directed toward the problem of machine translation, though this provide one of the best tests beds imaginable.

We feel the time is now ripe for applying the latest techniques in artificial intelligence and natural-language processing (e.g., various systems of knowledge represention and "non-monotonic" reasoning) to the problem of translation. Further, it is imperative that research be conducted into problems specific to translation, such as the acquisition and representation of contrastive interlinguistic knowledge.

Our program therefore provides for the exploration of advance machine translation models (those belonging to what is called the "third generation" of MT systems) and the demonstration of such models in prototype systems.

CONCLUSION

Translation brings into play the whole range of human cognitive skills. For this reason, the universal translation machine remains beyond our grasp for the time being. There is still help that can be provided to translators whose workload is constantly increasing, to the extent that we adopt appropriate sub-optimal strategies: i.e., lower levels of automation or system specialization. In view of the renewed international interest in automated translation, action is urgently called for in Canada.

The aim of the CWARC's program is to help restore Canada to its traditional place as a leader in the field. For the short and medium term, the program takes a realistic approach, based on clearly defined sub-optimal strategies: the translator's workstation and systems based on sublanguages. In the longer term, its aim is to put artificial intelligence to work in the service of translation.

Notes

1. Bar-Hillel Y., The Present Status of Automatic Translation of Languages, in F.L. Alt. (ed.), Advances in Computers, vol. 1, Academic Press, New York, 1960, pp. 91-163.

2. Kay M., The Proper Place of Man and Machines in Translation, AJCL, 1976, microfiche 46.

3. Kittredge R., Lehrberger J. (eds.), Sublanguage: Studies of Language in Restricted Semantic Domains, De Gruyter, New York, 1982.

4. Isabelle, P., "The TAUM machine translation project", **La Revue québécoise de Linguistique**, 5:2, 1986.

5. ALPAC. Languages and Machines: Computers in Translation and Linguistics. Report of the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, NCR, publication 1416, Washington, 1966.

6. **Machine Translation and Natural-Language Processing:** Opportunities for Artificial Intelligence in Canada. Report of a study by the firm Cognos under contract from the Department of Communications, ref: DOC-CR-84-048, 1985.